Conditional Akaike information under covariate shift with application to small area prediction

Yuki Kawakubo* Shonosuke Sugasawa† and Tatsuya Kubokawa‡

Abstract

In this study, we consider the problem of selecting explanatory variables of fixed effects in linear mixed models under covariate shift, which is when the values of covariates in the predictive model differ from those in the observed model. We construct a variable selection criterion based on the conditional Akaike information introduced by Vaida and Blanchard (2005). We focus especially on covariate shift in small area prediction and demonstrate the usefulness of the proposed criterion. In addition, numerical performance is investigated through simulations, one of which is a design-based simulation using a real dataset of land prices.

Keywords: Akaike information criterion; Conditional AIC; Covariate shift; Linear mixed model; Small area estimation; Variable selection.

1 Introduction

Linear mixed models have been studied for a long time theoretically, and have many applications, for example, longitudinal data analysis in biostatistics, panel data analysis in econometrics, and small area estimation in official statistics. The problem of selecting explanatory variables in linear mixed models is important and has been investigated in the literature. Müller et al. (2013) is a good survey on model selection in linear mixed models.

When the purpose of the variable selection is to find a set of significant variables for a good prediction, Akaike-type information criteria (Akaike, 1973, 1974) are well-known methods. However, the Akaike information criterion (AIC) based on marginal likelihood, which integrates out likelihood with respect to random effects, is not appropriate when the prediction is focused on random effects. Then, Vaida and Blanchard (2005) proposed considering Akaike-type information based on the conditional density given the random effects and proposed the conditional AIC (cAIC). Here, we provide a brief explanation of the cAIC concept. Let $f(\boldsymbol{y}|\boldsymbol{b},\boldsymbol{\eta})$ be a conditional density function of $\boldsymbol{y}$ given $\boldsymbol{b}$, where $\boldsymbol{y}$ is an observable random vector of the response variables, $\boldsymbol{\eta}$ is a vector of the unknown parameters, and $\boldsymbol{b}$ is a random vector of the random effects. Let $\pi(\boldsymbol{b}|\boldsymbol{\eta})$ be a density function of $\boldsymbol{b}$. Then, Vaida and Blanchard (2005) proposed measuring the prediction risk of the plug-in predictive density $f(\tilde{\boldsymbol{y}}|\hat{\boldsymbol{b}},\widehat{\boldsymbol{\eta}})$ relative to Kullback–Leibler divergence:

$$\iint \left[\int \log \left\{ \frac{f(\tilde{\boldsymbol{y}}|\boldsymbol{b},\boldsymbol{\eta})}{f(\tilde{\boldsymbol{y}}|\hat{\boldsymbol{b}},\widehat{\boldsymbol{\eta}})} \right\} f(\tilde{\boldsymbol{y}}|\boldsymbol{b},\boldsymbol{\eta}) \mathrm{d}\tilde{\boldsymbol{y}} \right] f(\boldsymbol{y}|\boldsymbol{b},\boldsymbol{\eta}) \pi(\boldsymbol{b}|\boldsymbol{\eta}) \mathrm{d}\boldsymbol{y} \mathrm{d}\boldsymbol{b}, \tag{1}$$

*Faculty of Law, Politics and Economics, Chiba University (E-mail: kawakubo@chiba-u.jp)
†Institute of Statistical Mathematics
‡Faculty of Economics, University of Tokyo

where $\tilde{\boldsymbol{y}}$ is an independent replication of $\boldsymbol{y}$ given $\boldsymbol{b}$, and $\hat{\boldsymbol{b}}$ and $\widehat{\boldsymbol{\eta}}$ are some predictor or estimator of $\boldsymbol{b}$ and $\boldsymbol{\eta}$, respectively. The cAIC is an (asymptotically) unbiased estimator of a part of the risk in (1), which is called the conditional Akaike information (cAI), given by

$$\mathrm{cAI} = -2 \iiint \log \left\{ f(\tilde{\boldsymbol{y}}|\hat{\boldsymbol{b}}, \widehat{\boldsymbol{\eta}}) \right\} f(\tilde{\boldsymbol{y}}|\boldsymbol{b}, \boldsymbol{\eta}) f(\boldsymbol{y}|\boldsymbol{b}, \boldsymbol{\eta}) \pi(\boldsymbol{b}|\boldsymbol{\eta}) \mathrm{d}\tilde{\boldsymbol{y}} \mathrm{d}\boldsymbol{y} \mathrm{d}\boldsymbol{b}.$$

The cAIC as a variable selection criterion in linear mixed models has been studied by Liang et al. (2008), Greven and Kneib (2010), Srivastava and Kubokawa (2010), Kubokawa (2011), Kawakubo and Kubokawa (2014), and others. Furthermore, the cAIC has been constructed as a variable selection criterion in generalized linear mixed models by Donohue et al. (2011), Yu and Yau (2012), Yu et al. (2013), Saefken et al. (2014), and others.

Considering the prediction problem, it is often the case that the values of covariates in the predictive model are different from those in the observed model, which we call covariate shift. Here, we call the model in which $\boldsymbol{y}$ is the vector of the response variables the "observed model," and we call the model in which $\tilde{\boldsymbol{y}}$ is the vector of the response variables the "predictive model." It is noted that the term "covariate shift" was first used by Shimodaira (2000), who defined it as the situation in which the distribution of the covariates in the predictive model differs from that in the observed model. In this study, although we treat the covariates as non-random, we use the same term "covariate shift" as Shimodaira (2000). Even when the information about the covariates in the predictive model can be used, most of the Akaike-type information criteria do not use it. This is because it is assumed that the predictive model is the same as the observed model for deriving the criteria. As for the abovementioned cAIC, the conditional density of $\boldsymbol{y}$ given $\boldsymbol{b}$ and that of $\tilde{\boldsymbol{y}}$ given $\boldsymbol{b}$ are the same, both of which are denoted by $f(\cdot|\boldsymbol{b}, \boldsymbol{\eta})$. On the other hand, under the covariate shift, the conditional density of $\tilde{\boldsymbol{y}}$ given $\boldsymbol{b}$ is different from that of $\boldsymbol{y}$ given $\boldsymbol{b}$ and is denoted by $g(\tilde{\boldsymbol{y}}|\boldsymbol{b}, \boldsymbol{\eta})$. When the aim of the variable selection is to choose the best predictive model, it is not appropriate to use the covariates only in the observed model. Then, we redefine the cAI under covariate shift, as follows,

$$\mathrm{cAI} = -2 \iiint \log \left\{ g(\tilde{\boldsymbol{y}}|\hat{\boldsymbol{b}}, \widehat{\boldsymbol{\eta}}) \right\} g(\tilde{\boldsymbol{y}}|\boldsymbol{b}, \boldsymbol{\eta}) f(\boldsymbol{y}|\boldsymbol{b}, \boldsymbol{\eta}) \pi(\boldsymbol{b}|\boldsymbol{\eta}) \mathrm{d}\tilde{\boldsymbol{y}} \mathrm{d}\boldsymbol{y} \mathrm{d}\boldsymbol{b},$$

and construct an information criterion as an asymptotically unbiased estimator of the cAI. Satoh (1997, 2000) considered a similar problem in the multivariate linear regression model and proposed a variable selection criterion. It is important to note that we do not assume that the candidate model is overspecified, in other words, that the candidate model includes the true model. Although most of the Akaike-type information criteria make the overspecified assumption, this is not appropriate for estimating the cAI under covariate shift. We discuss this point in Section 3.2.

As an important applicable example of covariate shift, we focus on small area prediction, which is based on a finite super-population model. We consider the situation in which we are interested in the finite subpopulation (area) mean of some characteristic and that some values in the subpopulation are observed through some sampling procedure. When the sample size in each area is small, the problem is called small area estimation. For details about small area estimation, see Rao and Molina (2015), Datta and Ghosh (2012), Pfeffermann (2013), and others. The model-based approach in small area estimation often assumes that the finite population has a super-population with random effects and borrows strength from other areas to estimate (predict) the small area (finite subpopulation) mean. The well-known unit-level model is the nested error regression model (NERM), which is a kind of linear mixed model, and was introduced by Battese et al. (1988). The NERM can be used when the values of the auxiliary variables for

the units with characteristics of interest (response variables in the model) are observed through survey sampling. This is the observed model in the framework of our variable selection procedure. On the other hand, two types of predictive model can be considered. One is the unit-level model, which can be used in the situation in which the values of the auxiliary variables are known for all units. The other is the area-level model, which can be used in the situation in which each mean of the auxiliary variables is known for each small area. The latter is often the case in official statistics and the model introduced by Fay and Herriot (1979) is often used in this case.

The rest of this paper is organized as follows. In Section 2, we explain the setup of variable selection problem. In Section 3, we define the cAI under covariate shift in linear mixed models and obtain an asymptotically unbiased estimator of the cAI. In Section 4, we provide an example of covariate shift, which is focused on small area prediction. In Section 5, we investigate the numerical performance of the proposed criteria by simulations, one of which is design-based simulation based on a real dataset of land prices. All the proofs are given in the Appendix.

2 Setup of variable selection problem

2.1 Class of candidate models

We focus on the variable selection of the fixed effects. First, we consider the collection of candidate models as follows. Let $n \times p_\omega$ matrix $\boldsymbol{X}(\omega)$ consist of all the explanatory variables and assume that $\mathrm{rank}\,(\boldsymbol{X}(\omega)) = p_\omega$. In order to define candidate models by the index set j, suppose that j denotes a subset of $\omega = \{1, \ldots, p_\omega\}$ containing p_j elements, *i.e.*, $p_j = \#(j)$, and $\boldsymbol{X}(j)$ consists of p_j columns of $\boldsymbol{X}(\omega)$ indexed by the elements of j. We define the class of the candidate models by $\mathcal{J} = \mathcal{P}(\omega)$, namely, the power set of ω, in which we call ω the full model. We assume that the true model exists in the class of the candidate models $\mathcal{J}$, which is denoted by j_*. It is noteworthy that the dimension of the true model is p_{j_*}, which is abbreviated to p_*.

We next introduce the terms "overspecified" and "underspecified" models. Candidate model j is overspecified if $\boldsymbol{X}(\omega)\boldsymbol{\beta}_* \in \mathcal{R}[\boldsymbol{X}(j)]$, which means that $\boldsymbol{X}(\omega)\boldsymbol{\beta}_*$ is in the column space of $\boldsymbol{X}(j)$ following Fujikoshi and Satoh (1997) or Kawakubo and Kubokawa (2014). The set of overspecified models are denoted by $\mathcal{J}_+ = \{j \in \mathcal{J} | j_* \subseteq j\}$. On the other hand, candidate model j is underspecified when $\boldsymbol{X}(\omega) \nsubseteq \mathcal{R}[\boldsymbol{X}(j)]$. The set of underspecified models is denoted by $\mathcal{J}_- = \mathcal{J} \setminus \mathcal{J}_+$. It is important to note that most of the Akaike-type information criteria are derived under the assumption that the candidate model is overspecified. However, the assumption is not appropriate for considering the covariate shift, which is explained in Section 3.2. Thus, we derive the criterion without the overspecified assumption.

In the following two subsections, we clarify the observed model and predictive model for deriving the criteria.

2.2 Observed model

The candidate observed model j is the linear mixed model

$$\boldsymbol{y} = \boldsymbol{X}(j)\boldsymbol{\beta}_j + \boldsymbol{Z}\boldsymbol{b}_j + \boldsymbol{\varepsilon}_j, \tag{2}$$

where $\boldsymbol{y}$ is an $n \times 1$ observation vector of response variables, $\boldsymbol{X}(j)$ and $\boldsymbol{Z}$ are $n \times p_j$ and $n \times q$ matrixes of covariates, respectively, $\boldsymbol{\beta}_j$ is a $p_j \times 1$ vector of regression coefficients, $\boldsymbol{b}_j$ is a $q \times 1$ vector of random effects, and $\boldsymbol{\varepsilon}_j$ is an $n \times 1$ vector of random errors. Let $\boldsymbol{b}_j$ and $\boldsymbol{\varepsilon}_j$ be mutually

independent and $\boldsymbol{b}_j \sim \mathcal{N}_q(\boldsymbol{0}, \sigma_j^2 \boldsymbol{G})$, $\boldsymbol{\varepsilon}_j \sim \mathcal{N}_n(\boldsymbol{0}, \sigma_j^2 \boldsymbol{R})$, where $\boldsymbol{G}$ and $\boldsymbol{R}$ are $q \times q$ and $n \times n$ positive definite matrixes and σ_j^2 is a scalar. We assume that $\boldsymbol{G}$ and $\boldsymbol{R}$ are known and σ_j^2 is unknown.

The true observed model j_* is

$$\boldsymbol{y} = \boldsymbol{X}(\omega)\boldsymbol{\beta}_* + \boldsymbol{Z}\boldsymbol{b}_* + \boldsymbol{\varepsilon}_*,$$

where $\boldsymbol{b}_* \sim \mathcal{N}_q(\boldsymbol{0}, \sigma_*^2 \boldsymbol{G})$, $\boldsymbol{\varepsilon}_* \sim \mathcal{N}_n(\boldsymbol{0}, \sigma_*^2 \boldsymbol{R})$ and $\boldsymbol{\beta}_*$ is $p_\omega \times 1$ vector of regression coefficients, whose $p_\omega - p_*$ components are exactly 0 and the rest of the components are not 0. The fact that we can write the true model as the equation above implies that the true model exists in the class of candidate models $\mathcal{J}$. Note that $\boldsymbol{X}(\omega)$ is $n \times p_\omega$ matrix of covariates for the full model ω. Then, the marginal distribution of $\boldsymbol{y}$ is

$$\boldsymbol{y} \sim \mathcal{N}_n(\boldsymbol{X}(\omega)\boldsymbol{\beta}_*, \sigma_*^2 \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma} = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}} + \boldsymbol{R}$. For the true model, the conditional density function of $\boldsymbol{y}$ given $\boldsymbol{b}_*$ and the density function of $\boldsymbol{b}_*$ are denoted by $f(\boldsymbol{y}|\boldsymbol{b}_*, \boldsymbol{\beta}_*, \sigma_*^2)$ and $\pi(\boldsymbol{b}_*|\sigma_*^2)$, respectively.

2.3 Predictive model

The candidate predictive model j is the linear mixed model, which has the same regression coefficients $\boldsymbol{\beta}_j$ and random effects $\boldsymbol{b}_j$ as the candidate observed model j, but different covariates, namely,

$$\tilde{\boldsymbol{y}} = \widetilde{\boldsymbol{X}}(j)\boldsymbol{\beta}_j + \widetilde{\boldsymbol{Z}}\boldsymbol{b}_j + \tilde{\boldsymbol{\varepsilon}}_j, \tag{3}$$

where $\tilde{\boldsymbol{y}}$ is $m \times 1$ random vector of the target of prediction, $\widetilde{\boldsymbol{X}}(j)$ and $\widetilde{\boldsymbol{Z}}$ are $m \times p_j$ and $m \times q$ matrixes of covariates whose columns correspond to those of $\boldsymbol{X}(j)$ and $\boldsymbol{Z}$, respectively, and $\tilde{\boldsymbol{\varepsilon}}_j$ is $m \times 1$ vector of random errors, which is independent of $\boldsymbol{b}_j$ and $\boldsymbol{\varepsilon}_j$ and is distributed as $\tilde{\boldsymbol{\varepsilon}}_j \sim \mathcal{N}_m(\boldsymbol{0}, \sigma_j^2 \widetilde{\boldsymbol{R}})$, where $\widetilde{\boldsymbol{R}}$ is a known $m \times m$ positive definite matrix. We assume that we know the values of $\widetilde{\boldsymbol{X}}(j)$ and $\widetilde{\boldsymbol{Z}}$ in the predictive model and that they are not necessarily the same as those of $\boldsymbol{X}(j)$ and $\boldsymbol{Z}$ in the observed model. We call this situation covariate shift. The conditional density function of $\tilde{\boldsymbol{y}}$ given $\boldsymbol{b}_j$ for the model j is denoted by $g_j(\tilde{\boldsymbol{y}}|\boldsymbol{b}_j, \boldsymbol{\beta}_j, \sigma_j^2)$.

The true predictive model j_* is

$$\tilde{\boldsymbol{y}} = \widetilde{\boldsymbol{X}}(\omega)\boldsymbol{\beta}_* + \widetilde{\boldsymbol{Z}}\boldsymbol{b}_* + \tilde{\boldsymbol{\varepsilon}}_*,$$

where $\widetilde{\boldsymbol{X}}(\omega)$ is $m \times p_\omega$ matrix of covariates and $\tilde{\boldsymbol{\varepsilon}}_* \sim \mathcal{N}_m(\boldsymbol{0}, \sigma_*^2 \widetilde{\boldsymbol{R}})$. Then, the marginal distribution of $\tilde{\boldsymbol{y}}$ is

$$\tilde{\boldsymbol{y}} \sim \mathcal{N}_m(\widetilde{\boldsymbol{X}}(\omega)\boldsymbol{\beta}_*, \sigma_*^2 \widetilde{\boldsymbol{\Sigma}}),$$

where $\widetilde{\boldsymbol{\Sigma}} = \widetilde{\boldsymbol{Z}}\boldsymbol{G}\widetilde{\boldsymbol{Z}}^{\mathrm{T}} + \widetilde{\boldsymbol{R}}$. For the true model, the conditional density function of $\tilde{\boldsymbol{y}}$ given $\boldsymbol{b}_*$ is denoted by $g(\tilde{\boldsymbol{y}}|\boldsymbol{b}_*, \boldsymbol{\beta}_*, \sigma_*^2)$.

3 Proposed criteria

3.1 Conditional Akaike information under covariate shift

The cAI introduced by Vaida and Blanchard (2005) measures the prediction risk of the plug-in predictive density $g_j(\tilde{\boldsymbol{y}}|\hat{\boldsymbol{b}}_j, \widehat{\boldsymbol{\beta}}_j, \hat{\sigma}_j^2)$, where $\widehat{\boldsymbol{\beta}}_j$ and $\hat{\sigma}_j^2$ are maximum likelihood estimators of $\boldsymbol{\beta}_j$ and σ_j^2, respectively, which are given as

$$\widehat{\boldsymbol{\beta}}_j = (\boldsymbol{X}(j)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(j))^{-1}\boldsymbol{X}(j)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{y},$$
$$\hat{\sigma}_j^2 = (\boldsymbol{y} - \boldsymbol{X}(j)\widehat{\boldsymbol{\beta}}_j)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X}(j)\widehat{\boldsymbol{\beta}}_j)/n,$$

respectively, and $\hat{\boldsymbol{b}}_j$ is the empirical Bayes estimator of $\boldsymbol{b}_j$ for quadratic loss, which is given by

$$\hat{\boldsymbol{b}}_j = \boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X}(j)\widehat{\boldsymbol{\beta}}_j).$$

Then, the cAI under covariate shift is

$$\begin{aligned}\mathrm{cAI} &= E^{(\boldsymbol{y},\boldsymbol{b}_*)}E^{\tilde{\boldsymbol{y}}|\boldsymbol{b}_*}\left[-2\log\{g_j(\tilde{\boldsymbol{y}}|\hat{\boldsymbol{b}}_j,\widehat{\boldsymbol{\beta}}_j,\hat{\sigma}_j^2)\}\right] \\ &= E^{(\boldsymbol{y},\boldsymbol{b}_*)}E^{\tilde{\boldsymbol{y}}|\boldsymbol{b}_*}\left[m\log(2\pi\hat{\sigma}_j^2) + \log|\widetilde{\boldsymbol{R}}| + (\tilde{\boldsymbol{y}} - \widetilde{\boldsymbol{X}}(j)\widehat{\boldsymbol{\beta}}_j - \widetilde{\boldsymbol{Z}}\hat{\boldsymbol{b}}_j)^{\mathrm{T}}\widetilde{\boldsymbol{R}}^{-1}(\tilde{\boldsymbol{y}} - \widetilde{\boldsymbol{X}}(j))\widehat{\boldsymbol{\beta}}_j - \widetilde{\boldsymbol{Z}}\hat{\boldsymbol{b}}_j)/\hat{\sigma}_j^2\right],\end{aligned}$$

where $E^{(\boldsymbol{y},\boldsymbol{b}_*)}$ and $E^{\tilde{\boldsymbol{y}}|\boldsymbol{b}_*}$ denote expectation with respect to the joint distribution of $(\boldsymbol{y},\boldsymbol{b}_*) \sim f(\boldsymbol{y}|\boldsymbol{b}_*,\boldsymbol{\beta}_*,\sigma_*^2)\pi(\boldsymbol{b}_*|\sigma_*^2)$ and the conditional distribution of $\tilde{\boldsymbol{y}}$ given $\boldsymbol{b}_*$, namely $\tilde{\boldsymbol{y}}|\boldsymbol{b}_* \sim g(\tilde{\boldsymbol{y}}|\boldsymbol{b}_*,\boldsymbol{\beta}_*,\sigma_*^2)$, respectively. Taking expectation with respect to $\tilde{\boldsymbol{y}}|\boldsymbol{b}_* \sim g(\tilde{\boldsymbol{y}}|\boldsymbol{b}_*,\boldsymbol{\beta}_*,\sigma_*^2)$ and $\boldsymbol{b}_*|\boldsymbol{y} \sim \mathcal{N}_q(\tilde{\boldsymbol{b}}_*,\sigma_*^2(\boldsymbol{G} - \boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}\boldsymbol{G}))$ for $\tilde{\boldsymbol{b}}_* = \boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X}(\omega)\boldsymbol{\beta}_*)$, we obtain

$$\mathrm{cAI} = E\left[m\log(2\pi\hat{\sigma}_j^2) + \log|\widetilde{\boldsymbol{R}}| + \mathrm{tr}\,(\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{\Lambda})\cdot\sigma_*^2/\hat{\sigma}_j^2 + \boldsymbol{a}^{\mathrm{T}}\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{a}/\hat{\sigma}_j^2\right], \tag{4}$$

where

$$\begin{aligned}\boldsymbol{\Lambda} &= \widetilde{\boldsymbol{\Sigma}} - \widetilde{\boldsymbol{Z}}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}\boldsymbol{G}\widetilde{\boldsymbol{Z}}^{\mathrm{T}}, \\ \boldsymbol{a} &= (\widetilde{\boldsymbol{X}}(j)\widehat{\boldsymbol{\beta}}_j - \widetilde{\boldsymbol{X}}(\omega)\boldsymbol{\beta}_*) - \widetilde{\boldsymbol{Z}}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}(j)\widehat{\boldsymbol{\beta}}_j - \boldsymbol{X}(\omega)\boldsymbol{\beta}_*).\end{aligned} \tag{5}$$

3.2 Drawback of overspecified model assumption

Most of the Akaike-type information criteria are derived under the assumption that the candidate model includes the true model, namely, the overspecified assumption. Although the assumption seems to be too strong, the influence is restrictive in practice. This is because the likelihood part of the criterion is a naive estimator of the risk function, namely, the cAI in the context of the cAIC.

Under the covariate shift situation, however, we cannot construct the likelihood part as a good estimator of the cAI. That is, the drawback of overspecified assumption is more serious in the situation of covariate shift than the usual one. In Section 5.1, we show that an unbiased estimator of the cAI under the overspecified assumption $\widehat{\mathrm{cAI}}_u$ in (22) has large bias for estimating the cAI of the underspecified models.

Thus, we evaluate and estimate the cAI directly both for the overspecified models and underspecified models in the following subsection, which is essential work in selecting variables in covariate shift.

3.3 Evaluation and estimation of cAI

We evaluate the cAI in (4) both for the overspecified model and for the underspecified model. We assume that the full model ω is overspecified, that is, the collection of the overspecified models $\mathcal{J}_+$ is not an empty set. In addition, we assume that the size of the response variable in the predictive model m is of order $O(n)$.

When the candidate model j is overspecified, $n\hat{\sigma}_j^2/\sigma_*^2$ follows the chi-squared distribution. Then, we can evaluate the expectation in (4) exactly. However, for the underspecified model, this is not true. In this case, we asymptotically approximate the cAI as the following theorem.

Theorem 1 *For the overspecified case, it follows that* cAI $= E[m\log(2\pi\hat{\sigma}_j^2)] + \log|\widetilde{\boldsymbol{R}}| + R^*$, *where*

$$R^* = \frac{n\gamma}{n - p_j - 2},$$

for $\gamma = \mathrm{tr}\,(\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{\Lambda}) + \mathrm{tr}\,[\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{A}(\boldsymbol{X}(j)^{\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(j))^{-1}\boldsymbol{A}^{\top}]$ *and* $\boldsymbol{A} = \widetilde{\boldsymbol{X}}(j) - \widetilde{\boldsymbol{Z}}\boldsymbol{G}\boldsymbol{Z}^{\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(j)$. *For the underspecified case,* cAI *is approximated as*

$$\mathrm{cAI} = E[m\log(2\pi\hat{\sigma}_j^2)] + \log|\widetilde{\boldsymbol{R}}| + R^* + R_1 + R_2 + R_3 + R_4 + O(n^{-1}), \tag{6}$$

where

$$\begin{aligned} R_1 &= \gamma(\lambda - 1), \\ R_2 &= \gamma \cdot n^{-1}\{-2\lambda^3 + (p_j + 4)\lambda^2 - (p_j + 2)\}, \\ R_3 &= \lambda \cdot \boldsymbol{\beta}_*^{\top}\boldsymbol{B}^{\top}\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{B}\boldsymbol{\beta}_*/\sigma_*^2, \end{aligned}$$

and

$$R_4 = n^{-1}\{-2\lambda^3 + (p_j + 4)\lambda^2\} \times \boldsymbol{\beta}_*^{\top}\boldsymbol{B}^{\top}\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{B}\boldsymbol{\beta}_*/\sigma_*^2,$$

for $\lambda = 1/(1 + \delta)$,

$$\begin{aligned} \delta &= \boldsymbol{\beta}_*^{\top}\boldsymbol{X}(\omega)^{\top}(\boldsymbol{P}_\omega - \boldsymbol{P}_j)\boldsymbol{X}(\omega)\boldsymbol{\beta}_*/(n\sigma_*^2), \\ \boldsymbol{B} &= \widetilde{\boldsymbol{P}}_j\boldsymbol{X}(\omega) - \widetilde{\boldsymbol{X}}(\omega) + \widetilde{\boldsymbol{Z}}\boldsymbol{G}\boldsymbol{Z}^{\top}(\boldsymbol{P}_\omega - \boldsymbol{P}_j)\boldsymbol{X}(\omega), \\ \boldsymbol{P}_j &= \boldsymbol{\Sigma}^{-1}\boldsymbol{X}(j)(\boldsymbol{X}(j)^{\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(j))^{-1}\boldsymbol{X}(j)^{\top}\boldsymbol{\Sigma}^{-1}, \\ \boldsymbol{P}_\omega &= \boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\omega)(\boldsymbol{X}(\omega)^{\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\omega))^{-1}\boldsymbol{X}(\omega)^{\top}\boldsymbol{\Sigma}^{-1}, \end{aligned}$$

and

$$\widetilde{\boldsymbol{P}}_j = \widetilde{\boldsymbol{X}}(j)(\boldsymbol{X}(j)^{\top}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(j))^{-1}\boldsymbol{X}(j)^{\top}\boldsymbol{\Sigma}^{-1}.$$

When the candidate model j *is overspecified, it follows that* R_1, R_2, R_3, *and* R_4 *are exactly* 0.

Because the approximation of cAI in (6) includes unknown parameters, we have to provide an estimator of cAI for practical use. First, we obtain estimators of R_1 and R_2, which are polynomials of λ. We define $\hat{\lambda}$, $\widehat{\lambda^2}$, and $\widehat{\lambda^3}$ as

$$\begin{aligned} \hat{\lambda} &= \frac{n - p_j}{n - p_\omega}\frac{\hat{\sigma}_\omega^2}{\hat{\sigma}_j^2}, \\ \widehat{\lambda^2} &= \frac{(n - p_j)(n - p_j + 2)}{(n - p_\omega)(n - p_\omega + 2)}\left(\frac{\hat{\sigma}_\omega^2}{\hat{\sigma}_j^2}\right)^2, \end{aligned}$$

and

$$\widehat{\lambda^3} = \frac{(n - p_j)(n - p_j + 2)(n - p_j + 4)}{(n - p_\omega)(n - p_\omega + 2)(n - p_\omega + 4)}\left(\frac{\hat{\sigma}_\omega^2}{\hat{\sigma}_j^2}\right)^3.$$

When $j \in \mathcal{J}_+$, it follows that

$$\frac{\hat{\sigma}_\omega^2}{\hat{\sigma}_j^2} \sim \mathrm{Be}\left(\frac{n - p_\omega}{2}, \frac{p_\omega - p_j}{2}\right),$$

where $\mathrm{Be}(\cdot, \cdot)$ denotes the beta distribution. This implies that $E(\hat{\lambda}) = E(\widehat{\lambda^2}) = E(\widehat{\lambda^3}) = 1$ for the overspecified case. For the underspecified case, on the other hand, it follows that $E[(\hat{\sigma}_\omega^2/\hat{\sigma}_j^2)^k] =$

$\lambda^k + O(n^{-1})$ as $n \to \infty$ for $k = 1, 2, 3$. Then, we obtain an estimator of R_2 in the approximation of cAI given by (6), which is given as follows:

$$\widehat{R_2} = \gamma \cdot \frac{-2\widehat{\lambda^3} + (p_j + 4)\widehat{\lambda^2} - (p_j + 2)}{n}. \tag{7}$$

Because R_1 is of order $O(n)$, we have to estimate λ with higher-order accuracy in order to obtain an estimator of R_1 whose bias is of order $O(n^{-1})$ for the underspecified case. To this end, we expand $E(\hat{\lambda})$ up to $O(n^{-1})$ as

$$\begin{aligned} E(\hat{\lambda}) &= \frac{n - p_j}{n - p_\omega} \cdot E\left(\frac{K_0}{K_0 + K_1}\right) \\ &= \lambda + \frac{-2\lambda^3 + (p_j + 2)\lambda^2 - p_j\lambda}{n} + O(n^{-2}). \end{aligned}$$

Then, we obtain an estimator of R_1 given as

$$\widehat{R_1} = \gamma \cdot \left\{ \hat{\lambda} - \frac{-2\widehat{\lambda^3} + (p_j + 2)\widehat{\lambda^2} - p_j\hat{\lambda}}{n} - 1 \right\}. \tag{8}$$

We now have the following lemma, which can be proved using Appendix C and Appendix D of Kawakubo and Kubokawa (2014).

Lemma 1 *When the candidate model j is underspecified, $\widehat{R_1}$ in (8) and $\widehat{R_2}$ in (7) are asymptotically unbiased estimators of R_1 and R_2, respectively, whose bias is of order $O(n^{-1})$, that is,*

$$E(\widehat{R_1}) = R_1 + O(n^{-1}) \quad \text{and} \quad E(\widehat{R_2}) = R_2 + O(n^{-1}).$$

When the candidate model j is overspecified, it follows that $E(\widehat{R_1}) = E(\widehat{R_2}) = 0$.

We next consider estimation procedures of R_3 and R_4, which are complex functions of unknown parameters. We observe R_3 and R_4 as functions of $\boldsymbol{\eta}_* = (\boldsymbol{\beta}_*^{\mathrm{T}}, \sigma_*^2)^{\mathrm{T}}$, that is, $R_3 = R_3(\boldsymbol{\eta}_*)$, $R_4 = R_4(\boldsymbol{\eta}_*)$ and substitute their unbiased estimators $\widetilde{\boldsymbol{\eta}} = (\widetilde{\boldsymbol{\beta}}^{\mathrm{T}}, \tilde{\sigma}^2)^{\mathrm{T}}$, which are given by

$$\begin{aligned} \widetilde{\boldsymbol{\beta}} &= \widehat{\boldsymbol{\beta}}_\omega = (\boldsymbol{X}(\omega)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\omega))^{-1}\boldsymbol{X}(\omega)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{y}, \\ \tilde{\sigma}^2 &= (\boldsymbol{y} - \boldsymbol{X}(\omega)\widetilde{\boldsymbol{\beta}})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X}(\omega)\widetilde{\boldsymbol{\beta}})/(n - p_\omega). \end{aligned}$$

Then, the plug-in estimators of R_3 and R_4 are

$$\begin{aligned} \widetilde{R_3} &= R_3(\widetilde{\boldsymbol{\eta}}) = \tilde{\lambda} \cdot \widetilde{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{B}^{\mathrm{T}}\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{B}\widetilde{\boldsymbol{\beta}}/\tilde{\sigma}^2, \\ \widetilde{R_4} &= R_4(\widetilde{\boldsymbol{\eta}}) = n^{-1}\{-2\tilde{\lambda}^3 + (p_j + 4)\tilde{\lambda}^2\} \times \widetilde{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{B}^{\mathrm{T}}\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{B}\widetilde{\boldsymbol{\beta}}/\tilde{\sigma}^2, \end{aligned} \tag{9}$$

where $\tilde{\lambda} = 1/(1 + \tilde{\delta})$ for

$$\tilde{\delta} = \widetilde{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{X}(\omega)^{\mathrm{T}}(\boldsymbol{P}_\omega - \boldsymbol{P}_j)\boldsymbol{X}(\omega)\widetilde{\boldsymbol{\beta}}/(n\tilde{\sigma}^2).$$

Because R_3 is of order $O(n)$, the plug-in estimator $\widetilde{R_3}$ has second-order bias. Then, we correct the bias by analytical method based on Taylor series expansions. We observe that expectation of the plug-in estimator $R_3(\widetilde{\boldsymbol{\eta}})$ is expanded as

$$E[R_3(\widetilde{\boldsymbol{\eta}})] = R_3(\boldsymbol{\eta}_*) + B_1(\boldsymbol{\eta}_*) + B_2(\boldsymbol{\eta}_*) + O(n^{-2}), \tag{10}$$

where $B_1(\boldsymbol{\eta}_*)$ is second-order bias and $B_2(\boldsymbol{\eta}_*)$ is third-order bias of $R_3(\widetilde{\boldsymbol{\eta}})$, that is, $B_1(\boldsymbol{\eta}_*) = O(1)$ and $B_2(\boldsymbol{\eta}_*) = O(n^{-1})$, respectively. Because $\widetilde{\boldsymbol{\beta}}$ and $\tilde{\sigma}^2$ are independent, it follows that

$$B_1(\boldsymbol{\eta}_*) = \frac{1}{2} \cdot \mathrm{tr}\left[\frac{\partial^2 R_3(\boldsymbol{\eta}_*)}{\partial \boldsymbol{\beta}_* \partial \boldsymbol{\beta}_*^{\mathrm{T}}} E[(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_*)(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_*)^{\mathrm{T}}]\right] + \frac{1}{2}\frac{\partial^2 R_3(\boldsymbol{\eta}_*)}{(\partial \sigma_*^2)^2} E[(\tilde{\sigma}^2 - \sigma_*^2)^2], \tag{11}$$

where $E[(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_*)(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_*)^{\mathrm{T}}] = \sigma_*^2(\boldsymbol{X}(\omega)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\omega))^{-1}$ and $E[(\tilde{\sigma}^2 - \sigma_*^2)^2] = 2(\sigma_*^2)^2/(n - p_\omega)$. Second-order partial derivatives of R_3 are given by the following lemma.

Lemma 2 *The second-order partial derivative of $R_3(\boldsymbol{\eta}_*)$ with respect to $\boldsymbol{\beta}_*$ is*

$$\begin{aligned}\frac{\partial^2 R_3(\boldsymbol{\eta}_*)}{\partial \boldsymbol{\beta}_* \partial \boldsymbol{\beta}_*^{\mathrm{T}}} = \frac{\boldsymbol{\beta}_*^{\mathrm{T}}\boldsymbol{B}^{\mathrm{T}}\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{B}\boldsymbol{\beta}_*}{\sigma_*^2} \times \left\{-\frac{2\boldsymbol{C}}{n\sigma_*^2(1+\delta)^2} + \frac{8\boldsymbol{C}\boldsymbol{\beta}_*\boldsymbol{\beta}_*^{\mathrm{T}}\boldsymbol{C}}{n^2(\sigma_*^2)^2(1+\delta)^3}\right\} \\ - \frac{4\boldsymbol{B}^{\mathrm{T}}\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{B}\boldsymbol{\beta}_*\boldsymbol{\beta}_*^{\mathrm{T}}\boldsymbol{C} + 4\boldsymbol{C}\boldsymbol{\beta}_*\boldsymbol{\beta}_*^{\mathrm{T}}\boldsymbol{B}^{\mathrm{T}}\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{B}}{n(\sigma_*^2)^2(1+\delta)^2} + 2\lambda \cdot \boldsymbol{B}^{\mathrm{T}}\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{B}/\sigma_*^2,\end{aligned}$$

where $\boldsymbol{C} = \boldsymbol{X}(\omega)^{\mathrm{T}}(\boldsymbol{P}_\omega - \boldsymbol{P}_j)\boldsymbol{X}(\omega)$. The second-order partial derivative of $R_3(\boldsymbol{\eta}_*)$ with respect to σ_*^2 is

$$\begin{aligned}\frac{\partial^2 R_3(\boldsymbol{\eta}_*)}{(\partial \sigma_*^2)^2} = \frac{\boldsymbol{\beta}_*^{\mathrm{T}}\boldsymbol{B}^{\mathrm{T}}\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{B}\boldsymbol{\beta}_*}{\sigma_*^2} \times \left\{-\frac{2\boldsymbol{\beta}_*^{\mathrm{T}}\boldsymbol{C}\boldsymbol{\beta}_*}{n(\sigma_*^2)^3(1+\delta)^2} + \frac{2(\boldsymbol{\beta}_*^{\mathrm{T}}\boldsymbol{C}\boldsymbol{\beta}_*)^2}{n^2(\sigma_*^2)^4(1+\delta)^3}\right\} \\ - \frac{2\boldsymbol{\beta}_*^{\mathrm{T}}\boldsymbol{B}^{\mathrm{T}}\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{B}\boldsymbol{\beta}_*\boldsymbol{\beta}_*^{\mathrm{T}}\boldsymbol{C}\boldsymbol{\beta}_*}{n(\sigma_*^2)^4(1+\delta)^2} + 2\lambda \cdot \boldsymbol{\beta}_*^{\mathrm{T}}\boldsymbol{B}^{\mathrm{T}}\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{B}\boldsymbol{\beta}_*/(\sigma_*^2)^3.\end{aligned}$$

When the candidate model j is overspecified, second-order bias $B_1(\boldsymbol{\eta}_*)$ can be simplified to

$$B_1(\boldsymbol{\eta}_*) = \mathrm{tr}\,[\boldsymbol{B}^{\mathrm{T}}\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{B}(\boldsymbol{X}(\omega)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\omega))^{-1}],$$

because $\boldsymbol{C}\boldsymbol{\beta}_* = \boldsymbol{B}\boldsymbol{\beta}_* = \boldsymbol{0}$ and $\lambda = 1$, which implies that $(\partial^2 R_3(\boldsymbol{\eta}_*))/(\partial \boldsymbol{\beta}_* \partial \boldsymbol{\beta}_*^{\mathrm{T}}) = 2\boldsymbol{B}^{\mathrm{T}}\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{B}/\sigma_*^2$ and that $(\partial^2 R_3(\boldsymbol{\eta}_*))/(\partial \sigma_*^2)^2 = 0$. However, we cannot know which candidate models are overspecified. Then, we propose the following bias-corrected estimator of R_3:

$$\widetilde{\widetilde{R_3}} = R_3(\widetilde{\boldsymbol{\eta}}) - B_1(\widetilde{\boldsymbol{\eta}}). \tag{12}$$

Lemma 3 *Both for the cases in which the candidate model j is overspecified and j is underspecified, $\widetilde{\widetilde{R_3}}$ and $\widetilde{R_4}$ in (12) and (9) are asymptotically unbiased estimators of R_3 and R_4, whose bias is of order $O(n^{-1})$, that is,*

$$E(\widetilde{\widetilde{R_3}}) = R_3 + O(n^{-1}), \quad \text{and} \quad E(\widetilde{R_4}) = R_4 + O(n^{-1}).$$

Using $\widehat{R_1}$, $\widehat{R_2}$, $\widetilde{\widetilde{R_3}}$, and $\widetilde{R_4}$ given by (8), (7), (12), and (9), respectively, we construct an estimator of cAI as follows:

$$\widehat{\mathrm{cAI}} = m\log(2\pi\hat{\sigma}_j^2) + \log|\widetilde{\boldsymbol{R}}| + R^* + \widehat{R_1} + \widehat{R_2} + \widetilde{\widetilde{R_3}} + \widetilde{R_4}. \tag{13}$$

Then, we obtain the following theorem, which is shown by Theorem 1 and Lemmas 1–3.

Theorem 2 *Both for the cases in which the candidate model j is overspecified and j is underspecified, $\widehat{\mathrm{cAI}}$ in* (13) *is a second-order asymptotically unbiased estimator of* cAI*, that is,*

$$E(\widehat{\mathrm{cAI}}) = \mathrm{cAI} + O(n^{-1}).$$

When the sample size n is small, second-order accuracy seems not to be sufficient for the overspecified model. Actually, as the result in the simulation study shows, the estimate of the cAI of the true model has relatively large bias, although the estimation of the true model is important. Moreover, some of the other information criteria, which include the cAIC of Vaida and Blanchard (2005), are exact unbiased estimators of the information of the overspecified candidate model. Thus, we should improve the estimators of R_3 and R_4 to remove the bias that is of order $O(n^{-1})$. To this end, we adopt a parametric bootstrap method.

Bootstrap sample $\boldsymbol{y}^\dagger$ is generated by

$$\boldsymbol{y}^\dagger = \boldsymbol{X}(\omega)\widetilde{\boldsymbol{\beta}} + \boldsymbol{Z}\boldsymbol{b}^\dagger + \boldsymbol{\varepsilon}^\dagger,$$

where $\boldsymbol{b}^\dagger$ and $\boldsymbol{\varepsilon}^\dagger$ are generated by the following distribution:

$$\boldsymbol{b}^\dagger \sim \mathcal{N}(\boldsymbol{0}, \tilde{\sigma}^2\boldsymbol{G}), \quad \text{and} \quad \boldsymbol{\varepsilon}^\dagger \sim \mathcal{N}(\boldsymbol{0}, \tilde{\sigma}^2\boldsymbol{I}_n).$$

Then, we use the following estimator of R_4:

$$\widehat{R_4} = 2R_4(\widetilde{\boldsymbol{\eta}}) - E_{\widetilde{\boldsymbol{\eta}}}[R_4(\widetilde{\boldsymbol{\eta}}^\dagger)] \tag{14}$$

where $E_{\widetilde{\boldsymbol{\eta}}}$ denotes expectation with respect to the bootstrap distribution and $\widetilde{\boldsymbol{\eta}}^\dagger$ is $\widetilde{\boldsymbol{\eta}}^\dagger = ((\widetilde{\boldsymbol{\beta}}^\dagger)^{\mathrm{T}}, \tilde{\sigma}^{2\dagger})^{\mathrm{T}}$ for

$$\begin{aligned}
\widetilde{\boldsymbol{\beta}}^\dagger &= (\boldsymbol{X}(\omega)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\omega))^{-1}\boldsymbol{X}(\omega)\boldsymbol{\Sigma}^{-1}\boldsymbol{y}^\dagger, \\
\tilde{\sigma}^{2\dagger} &= (\boldsymbol{y}^\dagger - \boldsymbol{X}(\omega)\widetilde{\boldsymbol{\beta}}^\dagger)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}^\dagger - \boldsymbol{X}(\omega)\widetilde{\boldsymbol{\beta}}^\dagger)/(n - p_\omega).
\end{aligned}$$

As for R_3, it follows from (10) that

$$E_{\widetilde{\boldsymbol{\eta}}}[R_3(\widetilde{\boldsymbol{\eta}}^\dagger)] = R_3(\widetilde{\boldsymbol{\eta}}) + B_1(\widetilde{\boldsymbol{\eta}}) + B_2(\widetilde{\boldsymbol{\eta}}) + O_p(n^{-2}).$$

However, $B_1(\widetilde{\boldsymbol{\eta}})$ has a bias with order $O(n^{-1})$, that is,

$$E[B_1(\widetilde{\boldsymbol{\eta}})] = B_1(\boldsymbol{\eta}_*) + B_{11}(\boldsymbol{\eta}_*) + O(n^{-2}),$$

where $B_{11}(\boldsymbol{\eta}_*) = O(n^{-1})$. Because this bias is not negligible when we want to estimate R_3 with third-order accuracy, we estimate the bias by bootstrap method as follows:

$$\widehat{B_{11}} = E_{\widetilde{\boldsymbol{\eta}}}[B_1(\widetilde{\boldsymbol{\eta}}^\dagger)] - B_1(\widetilde{\boldsymbol{\eta}}).$$

Then, we obtain an estimator of R_3, which is given as

$$\begin{aligned}
\widehat{R_3} &= 2R_3(\widetilde{\boldsymbol{\eta}}) - E_{\widetilde{\boldsymbol{\eta}}}[R_3(\widetilde{\boldsymbol{\eta}}^\dagger)] + \widehat{B_{11}} \\
&= 2R_3(\widetilde{\boldsymbol{\eta}}) - E_{\widetilde{\boldsymbol{\eta}}}[R_3(\widetilde{\boldsymbol{\eta}}^\dagger)] + E_{\widetilde{\boldsymbol{\eta}}}[B_1(\widetilde{\boldsymbol{\eta}}^\dagger)] - B_1(\widetilde{\boldsymbol{\eta}}).
\end{aligned} \tag{15}$$

Lemma 4 *Both for the cases in which the candidate model j is overspecified and j is underspecified, $\widehat{R_3}$ and $\widehat{R_4}$ in* (15) *and* (14) *are asymptotically unbiased estimators of R_3 and R_4, whose bias is of order $O(n^{-2})$, that is,*

$$E(\widehat{R_3}) = R_3 + O(n^{-2}), \quad \text{and} \quad E(\widehat{R_4}) = R_4 + O(n^{-2}).$$

Using $\widehat{R_1}$, $\widehat{R_2}$, $\widehat{R_3}$, and $\widehat{R_4}$ given by (8), (7), (15), and (14), we obtain an estimator of cAI as follows:

$$\widehat{\mathrm{cAI}}^{\dagger} = m\log(2\pi\hat{\sigma}_j^2) + \log|\widetilde{\boldsymbol{R}}| + R^* + \widehat{R_1} + \widehat{R_2} + \widehat{R_3} + \widehat{R_4}, \tag{16}$$

which improves $\widehat{\mathrm{cAI}}$ in unbiasedness. Then, we obtain the following theorem, which is proved by Theorem 2 and Lemma 4.

Theorem 3 *When the candidate model j is overspecified, $\widehat{\mathrm{cAI}}^{\dagger}$ in* (16) *is a third-order asymptotically unbiased estimator of* cAI*, that is,*

$$E(\widehat{\mathrm{cAI}}^{\dagger}) = \mathrm{cAI} + O(n^{-2}).$$

When the candidate model j is underspecified, $\widehat{\mathrm{cAI}}^{\dagger}$ is a second-order asymptotically unbiased estimator of cAI*, that is,*

$$E(\widehat{\mathrm{cAI}}^{\dagger}) = \mathrm{cAI} + O(n^{-1}).$$

4 Application to small area prediction

A typical example of the covariate shift situation appears in the small area prediction problem. The model for small area prediction supposes that the observed small area data have a finite population, which has the super-population model with random effects, one of which is the well-known NERM proposed by Battese et al. (1988).

Let Y_{ik} and $\boldsymbol{x}_{ik}(j)$ denote the value of a characteristic of interest and its p_j-dimensional auxiliary variable for the kth unit of the ith area for $i = 1, \ldots, q$ and $k = 1, \ldots, N_i$. Note that $\boldsymbol{x}_{ik}(j)$ is a subvector of $\boldsymbol{x}_{ik}(\omega)$, which is the vector of the explanatory variables in the full model ω, and we hereafter abbreviate the model index j and write $\boldsymbol{x}_{ik}$ instead of $\boldsymbol{x}_{ik}(j)$, p instead of p_j, etc. Then, the NERM is

$$Y_{ik} = \boldsymbol{x}_{ik}^{\mathrm{T}}\boldsymbol{\beta} + b_i + \varepsilon_{ik} \quad (i = 1, \ldots, q;\ k = 1, \ldots, N_i), \tag{17}$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients, b_i is a random effect for the ith area, and b_is and ε_{ik}s are mutually independently distributed as $b_i \sim \mathcal{N}(0, \tau^2)$ and $\varepsilon_{ik} \sim \mathcal{N}(0, \sigma^2)$, respectively. We consider the situation in which only n_i values of the Y_{ik}s are observed through some sampling procedure. We define the number of unobserved variables in the ith area by $N_i - n_i = r_i$ and let $n = n_1 + \cdots + n_q, r = r_1 + \cdots + r_q$. Suppose, without loss of generality, the first n_i elements of $\{Y_{i1}, \ldots, Y_{i,N_i}\}$ are observed, which are denoted by $y_1, \ldots, y_{i,n_i}$, and $Y_{i,n_i+1}, \ldots, Y_{i,N_i}$ are unobserved. Then, the observed model is defined as

$$y_{ik} = \boldsymbol{x}_{ik}^{\mathrm{T}}\boldsymbol{\beta} + b_i + \varepsilon_{ik} \quad (i = 1, \ldots, q;\ k = 1, \ldots, n_i), \tag{18}$$

which corresponds to (2) with $\boldsymbol{y} = (\boldsymbol{y}_1^{\mathrm{T}}, \ldots, \boldsymbol{y}_q^{\mathrm{T}})^{\mathrm{T}}$ for $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{i,n_i})^{\mathrm{T}}$, $\boldsymbol{X} = (\boldsymbol{X}_1^{\mathrm{T}}, \ldots, \boldsymbol{X}_q^{\mathrm{T}})^{\mathrm{T}}$ for $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{i,n_i})^{\mathrm{T}}$, $\boldsymbol{Z} = \mathrm{diag}\,(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_q)$ for $\boldsymbol{Z}_i = \boldsymbol{1}_{n_i}$, $\boldsymbol{G} = \psi\boldsymbol{I}_q$ and $\boldsymbol{R} = \boldsymbol{I}_n$, where $\boldsymbol{1}_{n_i}$ denotes an $n_i \times 1$ vector of 1s and $\psi = \tau^2/\sigma^2$. In the derivation of our proposed criteria, we assume that the covariance matrix of $\boldsymbol{b}$ is $\sigma^2\boldsymbol{G}$ for a known matrix $\boldsymbol{G}$. However, in the NERM, $\boldsymbol{G}$ includes the parameter ψ, which is usually unknown and has to be estimated. In this case, we propose that $\boldsymbol{G}$ in the bias correction should be replaced with its plug-in estimator $\boldsymbol{G}(\widehat{\psi})$. The influence caused by the replacement may be limited because ψ is the nuisance parameter when we are interested in selecting only explanatory variables. Kawakubo and Kubokawa (2014) discussed the problem in their Remark 3.1.

We consider two types of predictive models. The first can be used in the situation in which all $\boldsymbol{x}_{ik}$s are available. Then, the predictive model, which we call the "unit-level predictive model," is defined by

$$Y_{ik} = \boldsymbol{x}_{ik}^{\mathrm{T}}\boldsymbol{\beta} + b_i + \varepsilon_{ik} \quad (i = 1, \ldots, q;\ k = n_i + 1, \ldots, N_i), \tag{19}$$

which corresponds to (3) with $\tilde{\boldsymbol{y}} = (\tilde{\boldsymbol{y}}_1^{\mathrm{T}}, \ldots, \tilde{\boldsymbol{y}}_q^{\mathrm{T}})^{\mathrm{T}}$ for $\tilde{\boldsymbol{y}}_i = (Y_{i,n_i+1}, \ldots, Y_{i,N_i})^{\mathrm{T}}$, $\widetilde{\boldsymbol{X}} = (\widetilde{\boldsymbol{X}}_1^{\mathrm{T}}, \ldots, \widetilde{\boldsymbol{X}}_q^{\mathrm{T}})^{\mathrm{T}}$ for $\widetilde{\boldsymbol{X}}_i = (\boldsymbol{x}_{i,n_i+1}, \ldots, \boldsymbol{x}_{i,N_i}^{\mathrm{T}})^{\mathrm{T}}$, $\widetilde{\boldsymbol{Z}} = \mathrm{diag}\,(\widetilde{\boldsymbol{Z}}_1, \ldots, \widetilde{\boldsymbol{Z}}_q)$ for $\widetilde{\boldsymbol{Z}}_i = \boldsymbol{1}_{r_i}$, $\widetilde{\boldsymbol{R}} = \boldsymbol{I}_r$. Note that $m = r$.

In the problem of small area prediction, we often encounter the situation in which all $\boldsymbol{x}_{ik}$s are not observed but the area mean $\bar{\boldsymbol{x}}_i = N_i^{-1}\sum_{k=1}^{N_i}\boldsymbol{x}_{ik}$ is known and we are interested in predicting $\overline{Y}_i$, which is the mean of finite population $\{Y_{i1}, \ldots, Y_{i,N_i}\}$, by using the value of $\bar{\boldsymbol{x}}_i$. Then, the second type of predictive model, which we call the "area-level predictive model," is defined as

$$\overline{Y}_{i(u)} = \bar{\boldsymbol{x}}_{i(u)}^{\mathrm{T}}\boldsymbol{\beta} + b_i + \bar{\varepsilon}_{i(u)} \quad (i = 1, \ldots, q), \tag{20}$$

where $\overline{Y}_{i(u)} = r_i^{-1}\sum_{k=n_i+1}^{N_i} Y_{ik}$, the mean of unobserved variables, $\bar{\boldsymbol{x}}_{i(u)} = r_i^{-1}\sum_{k=n_i+1}^{N_i}\boldsymbol{x}_{ik}$, calculated from $\bar{\boldsymbol{x}}_i$ and $(\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i})$, and $\bar{\varepsilon}_{i(u)} = r_i^{-1}\sum_{k=n_i+1}^{N_i}\varepsilon_{ik}$ is distributed as $\mathcal{N}(0, \sigma^2/r_i)$. The model (20) corresponds to (3) with $\tilde{\boldsymbol{y}} = (\overline{Y}_{1(u)}, \ldots, \overline{Y}_{q(u)})^{\mathrm{T}}$, $\widetilde{\boldsymbol{X}} = (\bar{\boldsymbol{x}}_{1(u)}, \ldots, \bar{\boldsymbol{x}}_{q(u)})^{\mathrm{T}}$, $\widetilde{\boldsymbol{Z}} = \boldsymbol{I}_q$ and $\widetilde{\boldsymbol{R}} = \mathrm{diag}\,(\widetilde{R}_1, \ldots, \widetilde{R}_q)$ for $\widetilde{R}_i = 1/r_i$. Note that $m = q$.

After selecting explanatory variables with our proposed criteria, we predict $\overline{Y}_{i(u)}$ by the empirical best linear unbiased predictor $\widehat{\overline{Y}}_{i(u)} = \bar{\boldsymbol{x}}_{i(u)}^{\mathrm{T}}\widehat{\boldsymbol{\beta}} + \hat{b}_i$ and obtain a predictor of the finite population mean $\overline{Y}_i$, which is given as

$$\widehat{\overline{Y}}_i = \frac{1}{N_i}\left\{\sum_{k=1}^{n_i} y_{ik} + r_i\widehat{\overline{Y}}_{i(u)}\right\}. \tag{21}$$

Thus, covariate shift appears in standard models for small area prediction and the proposed criterion is important and useful in such a situation.

5 Simulations

5.1 Measuring the bias of estimating the true cAI by the criteria

In this subsection, we compare the performance of the criteria by measuring the bias of estimating the cAI. We consider a class of the nested candidate models $j_\alpha = \{1, \ldots, \alpha\}$ for $\alpha = 1, \ldots, p_\omega$ where $p_\omega = 7$. The true observed model is the NERM in (18) with $\sigma^2 = \tau^2 = 1$ and $n_i = r_i = 3$ for $i = 1, \ldots, q$. We consider the unit-level predictive model (19) for the first experiment and the area-level predictive model (20) for the second experiment. The explanatory variables in the full model $\boldsymbol{x}_{ik}(\omega)$'s $(i = 1, \ldots, q;\ k = 1, \ldots, N_i)$ are independently generated by $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$, where $\boldsymbol{\Sigma}_{\boldsymbol{x}} = 0.9\boldsymbol{I}_{p_\omega} + 0.1\boldsymbol{J}_{p_\omega}$ for $\boldsymbol{J}_{p_\omega} = \boldsymbol{1}_{p_\omega}\boldsymbol{1}_{p_\omega}^{\mathrm{T}}$. The true coefficient vector $\boldsymbol{\beta}_*$ is $\boldsymbol{\beta}_* = (\beta_1, \ldots, \beta_{p_*}, 0, 0)$ for $p_* = 5$ and β_ls $(1 \le l \le p_*)$ are generated by $\beta_l = 2 \times ((-1)^l/(l + 0.7)) \times U(1, 2)$ for a uniform random variable $U(1, 2)$ on the interval $(1, 2)$. The values of the explanatory variables $\boldsymbol{x}_{ik}$s and the vector of regression coefficients $\boldsymbol{\beta}_*$ are fixed through simulations.

For comparison, we consider the exact unbiased estimator, which is derived under the assumption that the candidate model is overspecified, given by

$$\widehat{\mathrm{cAI}}_u = m\log(2\pi\hat{\sigma}_j^2) + \log|\widetilde{\boldsymbol{R}}| + (\boldsymbol{y} - \boldsymbol{X}(j)\widehat{\boldsymbol{\beta}}_j - \boldsymbol{Z}\hat{\boldsymbol{b}}_j)^{\mathrm{T}}\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X}(j)\widehat{\boldsymbol{\beta}}_j - \boldsymbol{Z}\hat{\boldsymbol{b}}_j)/\hat{\sigma}_j^2 + \Delta_{\mathrm{CS}}, \tag{22}$$

where the bias correction term Δ_{CS} is

$$\Delta_{\mathrm{CS}} = \frac{n}{n-p_j-2}\left\{\mathrm{tr}\left[\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{\Lambda}\right] + \mathrm{tr}\left[\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{A}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{A}^{\mathrm{T}}\right]\right\} + \frac{n}{n-p_j}\left\{-\mathrm{tr}\left[\boldsymbol{R}\boldsymbol{\Sigma}^{-1}\right] + \mathrm{tr}\left[\boldsymbol{R}\boldsymbol{P}_j\right]\right\}.$$

Table 1: Relative bias of estimating cAI by $\widehat{\mathrm{cAI}}_u$, $\widehat{\mathrm{cAI}}$ and $\widehat{\mathrm{cAI}}^{\dagger}$ for unit-level predictive model

	True value	Relative bias		
Model	cAI	$\widehat{\mathrm{cAI}}_u$	$\widehat{\mathrm{cAI}}$	$\widehat{\mathrm{cAI}}^{\dagger}$
Pattern (a): $q = 10$				
j_1	206.38	-33.439	-0.33371	-0.080972
j_2	152.81	-18.840	-0.2414	-0.16414
j_3	140.79	-18.556	-0.23026	-0.46789
j_4	132.61	-11.451	0.26445	-0.19514
j_5	116.51	-0.0019291	1.5979	0.49514
j_6	122.46	-0.050686	0.77756	0.15429
j_7	128.88	0.0086256	0.09468	0.09468
Pattern (b): $q = 15$				
j_1	233.35	-0.22534	0.11255	0.1159
j_2	189.54	9.4310	0.24145	0.21667
j_3	177.28	14.197	0.42246	0.37347
j_4	163.62	-0.76563	0.32597	0.02934
j_5	152.94	0.13627	0.8566	0.25115
j_6	156.73	0.068668	0.53817	0.15002
j_7	161.65	0.083897	0.015869	0.015869
Pattern (c): $q = 20$				
j_1	299.12	4.3775	0.084838	0.084654
j_2	252.60	6.1677	0.24072	0.23581
j_3	250.27	-5.1825	-0.016634	0.010911
j_4	208.25	2.6115	0.36013	0.23929
j_5	197.52	0.25682	0.53321	0.26101
j_6	200.38	0.24977	0.40855	0.25647
j_7	203.57	0.22713	0.21272	0.21272

Tables 1 and 2 report the true values of the cAI and relative bias of estimating the cAI by the criteria $\widehat{\mathrm{cAI}}$ in (13), $\widehat{\mathrm{cAI}}^{\dagger}$ in (16), and $\widehat{\mathrm{cAI}}_u$ in (22), for the experiment using the unit-level predictive model and for that using the area-level predictive model, respectively. We handle the cases in which the number of the areas $q = 10,\ 15,\ 20$. The true values of cAI in each candidate model are calculated based on (4) with 10,000 Monte Carlo iterations. The relative bias of estimating the cAI by the criteria is defined as

$$100 \times \frac{E[\mathrm{IC}] - \mathrm{cAI}}{\mathrm{cAI}},$$

Table 2: Relative bias of estimating cAI by $\widehat{\text{cAI}}_u$, $\widehat{\text{cAI}}$ and $\widehat{\text{cAI}}^{\dagger}$ for area-level predictive model

	True value	Relative bias		
Model	cAI	$\widehat{\text{cAI}}_u$	$\widehat{\text{cAI}}$	$\widehat{\text{cAI}}^{\dagger}$
Pattern (a): $q = 10$				
j_1	61.095	-10.429	-0.27157	-0.21359
j_2	46.635	5.4222	0.40292	-0.055653
j_3	50.744	-10.285	0.36735	-0.23857
j_4	47.323	-0.77484	1.3412	0.36198
j_5	45.735	-0.21108	2.2751	0.60073
j_6	49.452	-0.32466	0.82334	0.017969
j_7	52.805	-0.29278	-0.082743	-0.082743
Pattern (b): $q = 15$				
j_1	95.393	-5.4716	0.12331	0.13930
j_2	70.056	17.521	0.39905	0.36599
j_3	66.412	21.039	0.59611	0.53872
j_4	61.310	5.1740	0.58453	0.10853
j_5	60.532	0.23723	1.0853	0.25515
j_6	63.109	0.13276	0.50306	0.096935
j_7	65.379	0.16648	-0.0017143	-0.0017143
Pattern (c): $q = 20$				
j_1	98.841	21.017	0.18161	0.17565
j_2	94.83	10.059	0.31607	0.30894
j_3	88.346	5.6464	0.13835	0.11302
j_4	78.383	7.8734	0.46942	0.27593
j_5	77.794	0.18930	0.54202	0.17009
j_6	79.277	0.18908	0.41365	0.18534
j_7	81.216	0.15395	0.11783	0.11783

where IC = $\widehat{\mathrm{cAI}}$, $\widehat{\mathrm{cAI}}^{\dagger}$, $\widehat{\mathrm{cAI}}_u$, and expectation is computed based on 1,000 replications. The bootstrap sample size is 1,000 for obtaining $\widehat{\mathrm{cAI}}^{\dagger}$. From the tables, we observe the following facts. First, $\widehat{\mathrm{cAI}}_u$ has large bias for underspecified models, that is, j_1, j_2, j_3, and j_4, while the modified estimators of the cAI, $\widehat{\mathrm{cAI}}$ and $\widehat{\mathrm{cAI}}^{\dagger}$, have smaller bias for both overspecified and underspecified models. Second, $\widehat{\mathrm{cAI}}^{\dagger}$ can estimate the cAI more unbiasedly than $\widehat{\mathrm{cAI}}$ can for the case of small sample size because $\widehat{\mathrm{cAI}}^{\dagger}$ is a third-order asymptotically unbiased estimator of the cAI. In particular, the improvement is remarkable for the true model j_5, which is important for variable selection. However, the relative bias of $\widehat{\mathrm{cAI}}$, which is the second-order asymptotically unbiased estimator of the cAI, becomes smaller as the sample size becomes larger and the difference in performance between $\widehat{\mathrm{cAI}}$ and $\widehat{\mathrm{cAI}}^{\dagger}$ is not very significant.

5.2 Predicting finite population mean

In this subsection, we investigate the numerical performance of the small area prediction problem explained in Section 4. We conduct design-based simulation based on a real dataset. We use the posted land price data along the Keikyu train line, which connects the suburbs in Kanagawa prefecture to the Tokyo metropolitan area. This dataset was also used by Kawakubo and Kubokawa (2014), who studied modification of the cAIC.

We analyze the land price data in 2001 with covariates for 47 stations that we consider as small areas, and let $q = 47$. In the original sample, there are n_i sampled land spots for the i area, and the total sample size is $n = \sum_{i=1}^{q} n_i = 189$. We generate a synthetic population of size $N = 1,000$ by resampling with replacement from the original dataset using selection probabilities inversely proportional to sample weights. This method of making a synthetic population was also used by Chandra et al. (2012). Then, we select 200 independent random samples, each of size $n = 189$, from the fixed synthetic population by sampling from each area based on simple random sampling without replacement and with sample size of each area equal to that of original dataset n_i.

The characteristic of interest is the land price (Yen in hundreds of thousands) per m^2 of the kth spot in the ith area, denoted by P_{ik}, and the target is the mean of the land price in each area $\overline{P}_i = N_i^{-1} \sum_{k=1}^{N_i} P_{ik}$ for $i = 1, \ldots, q$, where N_i is the size of the ith area (subpopulation). As discussed in Section 4, we adopt model-based estimation of finite subpopulation mean $\overline{P}_i$ by using NERM. For selecting the explanatory variables in NERM, we use our proposed criterion $\widehat{\mathrm{cAI}}$ in (13) for comparison with the conventional cAIC by Vaida and Blanchard (2005). However, because the land price data are right-skewed, we undertake log-transformation, namely, $Y_{ik} = \log(P_{ik})$, and fit Y_{ik} with NERM in (17).

The dataset includes the following auxiliary variables. FAR_{ik} denotes the floor-area ratio of spot k in the ith area, TRN_i is the time it takes by train from station i to Tokyo station around 9:00 AM, DST_{ik} is the geographical distance from spot k to nearby station i and FOOT_{ik} denotes the time it takes on foot from spot k to nearby station i. As the candidate explanatory variables, we consider 7 variables FAR_{ik}, TRN_i, TRN_i^2, DST_{ik}, DST_{ik}^2, FOOT_{ik}, and FOOT_{ik}^2, which are denoted by $x_1, \ldots, x_7$, and $x_0 = 1$ denotes a constant term. Using the criteria, we select the best combination of these variables.

Based on the best model selected by the criteria, we obtain a predictor $\widehat{\overline{P}}_i$ of the finite subpopulation mean of the land price in the ith area. However, log-transformed variable Y_{ik} is used in the NERM, and thus, we have to modify the predictor (21). The best predictor of out-of-sample P_{ik} $(i = 1, \ldots, q;\ k = n_i + 1, \ldots, N_i)$ is the conditional expectation given the data $\boldsymbol{y}$, namely, $\widehat{P}_{ik}(\boldsymbol{y}, \boldsymbol{\theta}) = E[P_{ik} \mid \boldsymbol{y}] = E[\exp(Y_{ik}) \mid \boldsymbol{y}]$, where $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \tau^2, \sigma^2)^{\mathrm{T}}$. Because the

conditional mean and variance of Y_{ik} given $\boldsymbol{y}$, denoted by μ_{ik} and V_i, are

$$\mu_{ik} = \boldsymbol{x}_{ik}^{\mathrm{T}}\boldsymbol{\beta} + \frac{\tau^2}{\sigma^2 + n_i\tau^2}\sum_{k'=1}^{n_i}(y_{ik'} - \boldsymbol{x}_{ik'}^{\mathrm{T}}\boldsymbol{\beta}), \quad V_i = \sigma^2 + \frac{\tau^2\sigma^2}{\sigma^2 + n_i\tau^2},$$

it follows that

$$\widehat{P}_{ik}(\boldsymbol{y}, \boldsymbol{\theta}) = \exp(\mu_{ik} + V_i/2),$$

Substituting $\widehat{\boldsymbol{\theta}}$ with $\boldsymbol{\theta}$, we obtain the empirical best predictor (EBP):

$$\widehat{P}_{ik}^{\mathrm{EBP}} = \widehat{P}_{ik}(\boldsymbol{y}, \widehat{\boldsymbol{\theta}}),$$

where $\widehat{\boldsymbol{\theta}}$ is some estimator of $\boldsymbol{\theta}$. Then, we use the following predictor of $\overline{P}_i$:

$$\widehat{\overline{P}}_i = \frac{1}{N_i}\left\{\sum_{k=1}^{n_i} P_{ik} + \sum_{k=n_i+1}^{N_i} \widehat{P}_{ik}^{\mathrm{EBP}}\right\}.$$

As $\widehat{\boldsymbol{\theta}}$, we use unbiased estimators of τ^2 and σ^2 proposed by Prasad and Rao (1990) and the GLS estimator of $\boldsymbol{\beta}$.

We measure the performance of this design-based simulation by mean squared error (MSE) of the predictor,

$$\mathrm{MSE}_i = \frac{1}{200}\sum_{t=1}^{200}(\widehat{\overline{P}}_i^{(t)} - \overline{P}_i^{(t)})^2,$$

where $\overline{P}_i^{(t)}$ and $\widehat{\overline{P}}^{(t)}$ are ith finite subpopulation mean and its predictor of the tth sampled data. We construct $\widehat{\mathrm{cAI}}$s using the unit-level predictive model (19) and the area-level predictive model (20), and let $\mathrm{MSE}_i^{\mathrm{u}}$ and $\mathrm{MSE}_i^{\mathrm{a}}$ denote the corresponding MSEs. To compare the performance, we compute the ratio of MSEs as follows:

$$\frac{\mathrm{MSE}_i^{\mathrm{u}}}{\mathrm{MSE}_i^{\mathrm{vb}}}, \quad \text{and} \quad \frac{\mathrm{MSE}_i^{\mathrm{a}}}{\mathrm{MSE}_i^{\mathrm{vb}}}$$

where $\mathrm{MSE}_i^{\mathrm{vb}}$ is the MSE of the predictor based on the cAIC of Vaida and Blanchard (2005). Figure 1 shows the results. Although the performance of $\widehat{\mathrm{cAI}}$ based on the unit-level predictive model is similar to the cAIC of Vaida and Blanchard (2005), $\widehat{\mathrm{cAI}}$ based on the area-level predictive model has much better performances in most areas. It is valuable to point out that the MSE of the predictor of the finite subpopulation mean can be improved using our proposed criteria, which motivates us to use them for variable selection in the small area prediction problem.

Acknowledgments

We are grateful to Professor J.N.K. Rao for his helpful comment and for holding a seminar at Statistics Canada during our stay in Ottawa. This research was supported by Grant-in-Aid for Scientific Research from the Japan Society for the Promotion of Science, Grant Numbers 16K17101, 16H07406, 15H01943 and 26330036.

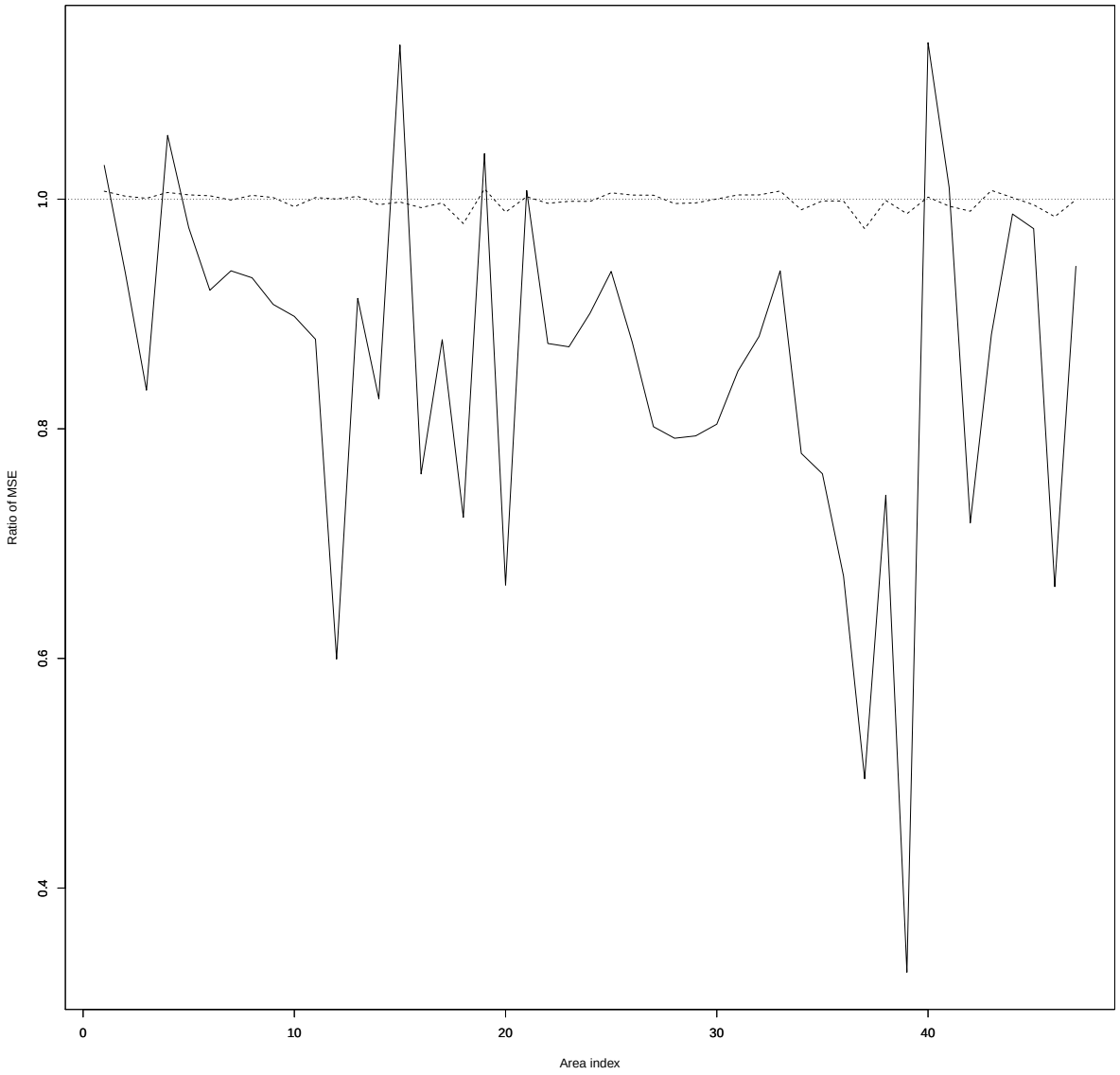

Figure 1: Ratios of MSE based on cAIC of Vaida and Blanchard (2005) to $\widehat{\text{cAI}}$ based on area-level predictive model (solid line) and to $\widehat{\text{cAI}}$ based on unit-level predictive model (dashed line)

A Appendix

A.1 Proof of Theorem 1

Because $\widehat{\boldsymbol{\beta}}_j$ and $\hat{\sigma}_j^2$ are mutually independent, the cAI in (4) can be rewritten as

$$\mathrm{cAI} = E[m\log(2\pi\hat{\sigma}_j^2)] + \log|\widetilde{\boldsymbol{R}}| + n\cdot E\big[(n\hat{\sigma}_j^2/\sigma_*^2)^{-1}\big]\big\{\mathrm{tr}\,(\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{\Lambda}) + E[\boldsymbol{a}^{\mathrm{T}}\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{a}/\sigma_*^2]\big\},$$

For the overspecified case, we can easily evaluate the cAI, noting that $n\hat{\sigma}_j^2/\sigma_*^2$ follows chi-squared distribution with $n-p_j$ degrees of freedom and is independent of $\widehat{\boldsymbol{\beta}}_j$. Thus, it suffices to show that the cAI is evaluated as (6) for the underspecified case.

From (B.4) in Kawakubo and Kubokawa (2014), we can evaluate $E[(n\hat{\sigma}_j^2/\sigma_*^2)^{-1}]$ as follows:

$$E[(n\hat{\sigma}_j^2/\sigma_*^2)^{-1}] = \frac{\lambda}{n}\left\{1+\frac{-2\lambda^2+(p_j+4)\lambda}{n}\right\} + O(n^{-3}). \tag{23}$$

We next evaluate $E[\boldsymbol{a}^{\mathrm{T}}\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{a}/\sigma_*^2]$. Let $\boldsymbol{u} = \boldsymbol{y}-\boldsymbol{X}(\omega)\boldsymbol{\beta}_*$. Then, we can rewrite $\widetilde{\boldsymbol{X}}(j)\widehat{\boldsymbol{\beta}}_j - \widetilde{\boldsymbol{X}}(\omega)\boldsymbol{\beta}_*$ in $\boldsymbol{a}$ as

$$\begin{aligned}\widetilde{\boldsymbol{X}}(j)\widehat{\boldsymbol{\beta}}_j - \widetilde{\boldsymbol{X}}(\omega)\boldsymbol{\beta}_* &= \widetilde{\boldsymbol{P}}_j(\boldsymbol{X}(\omega)\boldsymbol{\beta}_* + \boldsymbol{u}) - \widetilde{\boldsymbol{X}}(\omega)\boldsymbol{\beta}_* \\ &= (\widetilde{\boldsymbol{P}}_j\boldsymbol{X}(\omega) - \widetilde{\boldsymbol{X}}(\omega))\boldsymbol{\beta}_* + \sigma_*\widetilde{\boldsymbol{P}}_j\boldsymbol{\Sigma}^{1/2}\boldsymbol{v}.\end{aligned}$$

Next, we can rewrite $\boldsymbol{X}(j)\widehat{\boldsymbol{\beta}}_j - \boldsymbol{X}(\omega)\boldsymbol{\beta}_*$ in $\boldsymbol{a}$ as

$$\begin{aligned}\boldsymbol{X}(j)\widehat{\boldsymbol{\beta}}_j - \boldsymbol{X}(\omega)\boldsymbol{\beta}_* &= \sigma_*\boldsymbol{\Sigma}^{1/2}\{\boldsymbol{M}_j(\boldsymbol{W}_\omega\boldsymbol{\beta}_* + \boldsymbol{v}) - \boldsymbol{W}_\omega\boldsymbol{\beta}_*\} \\ &= \sigma_*\boldsymbol{\Sigma}^{1/2}\{-(\boldsymbol{M}_\omega - \boldsymbol{M}_j)\boldsymbol{W}_\omega\boldsymbol{\beta}_* + \boldsymbol{M}_j\boldsymbol{v}\} \\ &= -\boldsymbol{\Sigma}(\boldsymbol{P}_\omega - \boldsymbol{P}_j)\boldsymbol{X}(\omega)\boldsymbol{\beta}_* + \sigma_*\boldsymbol{\Sigma}^{1/2}\boldsymbol{M}_j\boldsymbol{v}.\end{aligned}$$

Then, we obtain

$$\boldsymbol{a} = \boldsymbol{B}\boldsymbol{\beta}_* + \sigma_*(\widetilde{\boldsymbol{P}}_j\boldsymbol{\Sigma}^{1/2} - \widetilde{\boldsymbol{Z}}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{M}_j)\boldsymbol{v}.$$

Moreover, it follows that

$$\begin{aligned}\widetilde{\boldsymbol{P}}_j\boldsymbol{\Sigma}^{1/2} - \widetilde{\boldsymbol{Z}}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{M}_j &= (\widetilde{\boldsymbol{X}}(j) - \widetilde{\boldsymbol{Z}}\boldsymbol{G}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(j))(\boldsymbol{X}(j)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(j))^{-1}\boldsymbol{X}(j)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2} \\ &= \boldsymbol{A}(\boldsymbol{X}(j)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(j))^{-1}\boldsymbol{X}(j)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\end{aligned}$$

Thus, $E[\boldsymbol{a}^{\mathrm{T}}\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{a}/\sigma_*^2]$ can be evaluated as

$$E[\boldsymbol{a}^{\mathrm{T}}\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{a}/\sigma_*^2] = \mathrm{tr}\,[\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{A}(\boldsymbol{X}(j)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}(j))^{-1}\boldsymbol{A}^{\mathrm{T}}] + \boldsymbol{\beta}_*^{\mathrm{T}}\boldsymbol{B}^{\mathrm{T}}\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{B}\boldsymbol{\beta}_*/\sigma_*^2. \tag{24}$$

It follows from (23) and (24) that

$$\begin{aligned}&n\cdot E[(K_0+K_1)^{-1}]\{\mathrm{tr}\,(\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{\Lambda}) + E[\boldsymbol{a}^{\mathrm{T}}\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{a}/\sigma_*^2]\} \\ &= (\gamma + \boldsymbol{\beta}_*^{\mathrm{T}}\boldsymbol{B}^{\mathrm{T}}\widetilde{\boldsymbol{R}}^{-1}\boldsymbol{B}\boldsymbol{\beta}_*/\sigma_*^2)\times\left\{\lambda + \frac{-2\lambda^3+(p_j+4)\lambda^2}{n}\right\} + O(n^{-1}),\end{aligned}$$

which shows that the cAI is approximated to (6). □

A.2 Proof of Lemma 2

First, note that

$$R_3(\boldsymbol{\eta}_*) = \lambda \cdot \boldsymbol{\beta}_*^{\mathrm{T}} \boldsymbol{B}^{\mathrm{T}} \widetilde{\boldsymbol{R}}^{-1} \boldsymbol{B} \boldsymbol{\beta}_* / \sigma_*^2,$$

where $\lambda = 1/(1+\delta)$ for

$$\delta = \boldsymbol{\beta}_*^{\mathrm{T}} \boldsymbol{X}(\omega)^{\mathrm{T}} (\boldsymbol{P}_\omega - \boldsymbol{P}_j) \boldsymbol{X}(\omega) \boldsymbol{\beta}_* / (n\sigma_*^2).$$

Then, we observe that

$$\frac{\partial R_3(\boldsymbol{\eta}_*)}{\partial \boldsymbol{\eta}_*} = \frac{\partial \lambda}{\partial \delta} \cdot \frac{\partial \delta}{\partial \boldsymbol{\eta}_*}$$

and that $\partial\lambda/\partial\delta = -(1+\delta)^{-2}$. After some calculations, we obtain Lemma 2. □

A.3 Proof of Lemma 3

First, note that $E[R_3(\widetilde{\boldsymbol{\eta}})]$ is expanded as

$$E[R_3(\widetilde{\boldsymbol{\eta}})] = R_3(\boldsymbol{\eta}_*) + B_1(\boldsymbol{\eta}_*) + O(n^{-1}),$$

where $B_1(\boldsymbol{\eta}_*)$ is given as (11). Because $B_1(\boldsymbol{\eta}_*) = O(1)$, it follows that $B_1(\widetilde{\boldsymbol{\eta}}) = B_1(\boldsymbol{\eta}_*) + O(n^{-1})$, which shows that

$$E[\widetilde{\widetilde{R_3}}] = R_3 + O(n^{-1}).$$

In the same way, we obtain $E[\widetilde{R_4}] = E[R_4(\widetilde{\boldsymbol{\eta}})] = R_4(\boldsymbol{\eta}_*) + O(n^{-1})$. □

A.4 Proof of Lemma 4

It follows from (15) that

$$E[\widehat{R_3}] = E\left[2R_3(\widetilde{\boldsymbol{\eta}}) - E_{\widetilde{\boldsymbol{\eta}}}[R_3(\widetilde{\boldsymbol{\eta}}^\dagger)] + E_{\widetilde{\boldsymbol{\eta}}}[B_1(\widetilde{\boldsymbol{\eta}}^\dagger)] - B_1(\widetilde{\boldsymbol{\eta}})\right].$$

Because $E[R_3(\widetilde{\boldsymbol{\eta}})]$ is expanded as $E[R_3(\widetilde{\boldsymbol{\eta}})] = R_3(\boldsymbol{\eta}_*) + B_1(\boldsymbol{\eta}_*) + B_2(\boldsymbol{\eta}_*) + O(n^{-2})$, we observe that

$$\begin{aligned} E\left[2R_3(\widetilde{\boldsymbol{\eta}}) - E_{\widetilde{\boldsymbol{\eta}}}[R_3(\widetilde{\boldsymbol{\eta}}^\dagger)]\right] &= 2\left\{R_3(\boldsymbol{\eta}_*) + B_1(\boldsymbol{\eta}_*) + B_2(\boldsymbol{\eta}_*)\right\} - E\left[R_3(\widetilde{\boldsymbol{\eta}}) + B_1(\widetilde{\boldsymbol{\eta}}) + B_2(\widetilde{\boldsymbol{\eta}})\right] + O(n^{-2}) \\ &= R_3(\boldsymbol{\eta}_*) + B_1(\boldsymbol{\eta}_*) + B_2(\boldsymbol{\eta}_*) - E\left[B_1(\widetilde{\boldsymbol{\eta}}) + B_2(\widetilde{\boldsymbol{\eta}})\right] + O(n^{-2}). \end{aligned}$$

Moreover, because $E[B_1(\widetilde{\boldsymbol{\eta}})] = B_1(\boldsymbol{\eta}_*) + B_{11}(\boldsymbol{\eta}_*) + O(n^{-2})$ and $E[B_2(\widetilde{\boldsymbol{\eta}})] = B_2(\boldsymbol{\eta}_*) + O(n^{-2})$, the equation above can be rewritten as

$$E\left[2R_3(\widetilde{\boldsymbol{\eta}}) - E_{\widetilde{\boldsymbol{\eta}}}[R_3(\widetilde{\boldsymbol{\eta}}^\dagger)]\right] = R_3(\boldsymbol{\eta}_*) - B_{11}(\boldsymbol{\eta}_*) + O(n^{-2}). \tag{25}$$

Next, it is observed that

$$\begin{aligned} E\left[E_{\widetilde{\boldsymbol{\eta}}}[B_1(\widetilde{\boldsymbol{\eta}}^\dagger)] - B_1(\widetilde{\boldsymbol{\eta}})\right] &= E[B_1(\widetilde{\boldsymbol{\eta}}) + B_{11}(\widetilde{\boldsymbol{\eta}})] - \{B_1(\boldsymbol{\eta}_*) + B_{11}(\boldsymbol{\eta}_*)\} + O(n^{-2}) \\ &= B_{11}(\boldsymbol{\eta}_*) + O(n^{-2}). \end{aligned} \tag{26}$$

Thus, it follows from (25) and (26) that

$$E[\widehat{R_3}] = R_3(\boldsymbol{\eta}_*) + O(n^{-2}).$$

Similarly, we show that $E[\widehat{R_4}] = R_4(\boldsymbol{\eta}_*) + O(n^{-2})$. □

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, (Petrov, B.N. and Csaki., F., eds.), 267–281, Akademia Kiado, Budapest.

Akaike, H. (1974). A new look at the statistical model identification. System identification and time-series analysis. *IEEE Transactions on Automatic Control*, **AC-19**, 716–723.

Battese, G.E., Harter, R.M., and Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, **83**, 28–36.

Chandra, H., Salvati, N., Chambers, R., and Tzavidis, N. (2012). Small area estimation under spatial nonstationarity. *Computational Statistics and Data Analysis*, **56**, 2875–2888.

Datta, G. and Ghosh, M. (2012). Small area shrinkage estimation. *Statistical Science*, **27**, 95–114.

Donohue, M.C., Overholser, R., Xu, R., and Vaida, F. (2011). Conditional Akaike information under generalized linear and proportional hazards mixed models. *Biometrika*, **98**, 685–700.

Fay, R.E. and Herriot, R.A. (1979). Estimates of income for small places: An application of James–Stein procedures to census data. *Journal of the American Statistical Association*, **74**, 269–277.

Fujikoshi, Y. and Satoh, K. (1997). Modified AIC and C_p in multivariate linear regression. *Biometrika*, **84**, 707–716.

Greven, S. and Kneib, T. (2010). On the behaviour of marginal and conditional AIC in linear mixed models. *Biometrika*, **97**, 773–789.

Kawakubo, Y. and Kubokawa, T. (2014). Modified conditional AIC in linear mixed models. *Journal of Multivariate Analysis*, **129**, 44–56.

Kubokawa, T. (2011). Conditional and unconditional methods for selecting variables in linear mixed model. *Journal of Multivariate Analysis*, **102**, 641–660.

Kubokawa, T. and Nagashima, B. (2012). Parametric bootstrap methods for bias correction in linear mixed models. *Journal of Multivariate Analysis*, **106**, 1–16.

Liang, H., Wu, H., and Zou, G. (2008). A note on conditional AIC for linear mixed-effects models. *Biometrika*, **95**, 773–778.

Müller, S., Scealy, J.L., and Welsh, A.H. (2013). Model selection in linear mixed models. *Statistical Science*, **28**, 135–167.

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*, **28**, 40–68.

Prasad, N.G.N. and Rao, J.N.K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, **85**, 163–171.

Rao, J.N.K. and Molina, I. (2015). *Small Area Estimation*, Wiley.

Saefken, B., Kneib, T., Van Waveren, C.S., and Greven, S. (2014). A unifying approach to the estimation of the conditional Akaike information in generalized linear mixed models. *Electronic Journal of Statistics*, **8**, 201–225.

Satoh, K. (1997). AIC-type model selection criterion for multivariate linear regression with a future experiment. *Journal of the Japan Statistical Society*, **27**, 135–140.

Satoh, K. (2000). Modification of AIC-type criterion in multivariate normal linear regression with a future experiment. *Hiroshima Mathematical Journal*, **30**, 29–53.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, **90**, 227–244.

Srivastava, M.S. and Kubokawa, T. (2010). Conditional information criteria for selecting variables in linear mixed models. *Journal of Multivariate Analysis*, **101**, 1970–1980.

Vaida, F. and Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika*, **92**, 351–370.

Yu, D. and Yau, K.K.W. (2012). Conditional Akaike information criterion of generalized linear mixed models. *Computational Statistics and Data Analysis*, **56**, 629–644.

Yu, D., Zhang, X., and Yau, K.K.W. (2013). Information based model selection criteria for generalized linear mixed models with unknown variance component parameters. *Journal of Multivariate Analysis*, **116**, 245–262.